Optimal Experimental Design in the Modelling of Pattern Formation

Adrián López García de Lomana, Àlex Gómez-Garrido,
David Sportouch, and Jordi Villà-Freixa

Grup de Recerca en Informàtica Biomèdica, IMIM-Universitat Pompeu Fabra,
C/Doctor Aiguader, 88, 08003 Barcelona, Catalunya, Spain
{adrianlopezgarciaadelomana,david.sportouch}@gmail.com,
{agomez,jvilla}@imim.es
<http://cbbl.imim.es>

Abstract. Gene regulation plays a major role in the control of developmental processes. Pattern formation, for example, is thought to be regulated by a limited number of genes translated into transcription factors that control the differential expression of other genes in different cells in a given tissue. We focused on the Notch pathway during the formation of chess-like patterns along development. Simplified models exist of the patterning by lateral inhibition due to the Notch-Delta signalling cascade. We show here how parameters from the literature are able to explain the steady-state behavior of model *tissues* of several sizes, although they are not able to reproduce time series of experiments. In order to refine the parameters set for data from real experiments we propose a practical implementation of an optimal experimental design protocol that combines parameter estimation tools with sensitivity analysis, in order to minimize the number of additional experiments to perform.

Keywords: lateral inhibition, GRN, optimal experimental design, multicellular system.

1 Introduction

One of the most breathtaking processes in biology is the development of a complex creature. In a matter of just a day (a fly maggot), a few weeks (a mouse) or several months (ourselves), an egg grows into millions, billions, or, in the case of humans, 10 trillion cells formed into organs, tissues and parts of the body. So, the main question in developmental biology is to understand how do cells arising from division of a single cell become different from each other. The complexity of the process of pattern formation in developmental biology has been dealt with by a number of researchers in the last decades (for reviews see [1]), both topologically, studying the different genes involved in the process and their relationships, and dynamically, measuring and modeling the temporal behavior of those genes and their products. Different simulation methods have been applied to dynamical models of patterning, involving both ordinary (ODE) [2]

and partial (PDE) differential equations or discrete representations of the cells as cellular automata, among others[3]. Initial models for pattern formation were based on simple assumptions that were able to capture most of the relevant information for a given general question. Thus, it is worth noting the efforts of Meinhardt [4] and others [5] in order to unravel the general rules governing the formation of complex patterns during the embryo development by using simple although soundable mathematical models.

At times, high throughput studies can be also performed in order to obtain time dependent qualitative information on the topology of the GRN. This type of information can be processed by probability and statistical inference tools that complement the verbal models defined by the experimentalists and provide a first formal model of the network. However, if one is able to quantify the dynamical information about the expression levels of different genes, even at the level of a few key genes by, for example, real time polymerase chain reaction (RT-PCR) experiments, global optimization protocols can be used to refine the parameters that describe the dynamics of the model.

In typical situations the modeller claims for experimental data that is scarce, low quality and, more importantly in most cases, difficult to obtain. How to maximize the outcome from limited resources is the aim of this paper. Here we present the implementation of a practical optimal experimental design pipeline in a theory/experiment integrated fashion. We demonstrate the utility of the protocol in the parameter estimation for one of the simplest models of pattern formation in biology, namely the Notch-Delta pathway for lateral inhibition (LI). To demonstrate the implementation of the method, we work with fictitious RT-PCR experimental data obtained from known models of the Notch-Delta interaction, as the LI model occurs between partner cells in a tissue, which offers an extra challenge for experimental manipulation. However, the proposed protocol is completely general for a RT-PCR experimental setting in any biological system that suits this technique.

2 Methods

2.1 Problem Statement

As outlined in the introduction, dynamical biological systems can be described by a large variety of mathematical models. Here we will restrict ourselves to models defined in terms of ODEs. Following [6], the time evolution of a system state of K species $\mathbf{x}(t, \theta) \in \mathbb{R}^K$ is solution of this set of ODEs:

$$\begin{cases} \frac{\partial}{\partial t} \mathbf{x}(t, \theta) = \mathbf{f}(\mathbf{x}(t, \theta), \theta, \mathbf{u}(t)). \\ \mathbf{x}(0) = \mathbf{x}_0. \end{cases} \quad (1)$$

Here $\theta \in \mathbb{R}^P$ denotes the parameters of the system, and $\mathbf{u}(t)$ is a vector containing the input of the system. The L properties of the system $\mathbf{y}^M(t, \theta) \in \mathbb{R}^{L_i}$ that can be measured are described by an observation function $\mathbf{g}$ at time t_i , $i = 1, \dots, N$ (N is here the number of design points):

$$\mathbf{y}^M(t_i, \boldsymbol{\theta}, u) = \mathbf{g}(\mathbf{x}(t_i, \boldsymbol{\theta}, \mathbf{u})), \quad i = 1, \dots, N. \quad (2)$$

The observations $\mathbf{Y}^D(t_i) \in \mathbb{R}^{L_i}$, $i = 1, \dots, N$ are considered as random variables and are given by

$$\mathbf{Y}^D(t_i) = \mathbf{y}^M(t_i, \boldsymbol{\theta}_0, \mathbf{u}) + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, N, \quad (3)$$

where $\boldsymbol{\theta}_0$ is the true parameter vector and $\boldsymbol{\epsilon}_i \in \mathbb{R}^{L_i}$, $i = 1, \dots, N$ describes the distribution error at time t_i . We assume that the distribution of the noise (observation error) follows a normal law (where the variances σ_{ij}^2 can be estimated from repetitions of the experiments):

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2), \quad j = 1, \dots, L_i, \quad i = 1, \dots, N. \quad (4)$$

In fact, $\mathbf{y}^M(t, \boldsymbol{\theta})$ refers to theoretical values (given by the model) and $\mathbf{y}^D(t_i)$ (realizations of the random variables $\mathbf{Y}^D(t_i)$) refers to practical values (it corresponds to the L_i measurements made experimentally at each time t_i , $i = 1, \dots, N$).

Maximum Likelihood Method. This method will help us to get estimates of the parameters of the system. In this method, we need to maximize the likelihood function $J_{ml}(\boldsymbol{\theta})$ to get estimates of the parameter vector $\boldsymbol{\theta}$. This function is defined as:

$$J_{ml}(\boldsymbol{\theta}; \mathbf{y}^D(t_1), \dots, \mathbf{y}^D(t_N)) = f(\mathbf{Y}^D(t_1), \dots, \mathbf{Y}^D(t_N))(\mathbf{y}^D(t_1), \dots, \mathbf{y}^D(t_N)). \quad (5)$$

As defined the random variables $\mathbf{Y}^D(t_i)$ follow a multivariate normal law

$$\mathbf{Y}^D(t_i) \sim \mathcal{N}_L(\mathbf{y}^M(t_i, \boldsymbol{\theta}, u), C(t_i)), \quad i = 1, \dots, N, \quad (6)$$

where $C(t_i) \in \mathcal{M}_{L_i, L_i}(\mathbb{R})$ is the covariance matrix defined as $C_{ll}(t_i) = \sigma_{il}^2$ and $C_{kl}(t_i) = 0$ if $k \neq l$.

$$J_{ml}(\boldsymbol{\theta}; \mathbf{y}^D(t_1), \dots, \mathbf{y}^D(t_N)) = \frac{1}{(2\pi)^{\frac{NL}{2}} \sqrt{\prod_{i=1}^N \prod_{j=1}^{L_i} \sigma_{ij}^2}} e^{-\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^{L_i} \left(\frac{y_j^D(t_i) - y_j^M(t_i, \boldsymbol{\theta}, u)}{\sigma_{ij}} \right)^2}. \quad (7)$$

Maximizing the likelihood function (regarding of $\boldsymbol{\theta}$) is in fact the same as maximizing the logarithm of the likelihood function, which in turn is the same as minimizing the opposite of the logarithm of the likelihood function. So, this leads to minimize the following function:

$$-ln(J_{ml}(\boldsymbol{\theta}; \mathbf{y}^D(t_1), \dots, \mathbf{y}^D(t_N))) = \frac{NL}{2} + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^{L_i} ln(\sigma_{ij}^2) + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^{L_i} \left(\frac{y_j^D(t_i) - y_j^M(t_i, \boldsymbol{\theta}, u)}{\sigma_{ij}} \right)^2. \quad (8)$$

So, this comes to minimize the following function:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{j=1}^{L_i} \left(\frac{y_j^D(t_i) - y_j^M(t_i, \boldsymbol{\theta}, u)}{\sigma_{ij}} \right)^2. \quad (9)$$

This corresponds to the minimization of a weighted residual sum of squares (with weights: $w_{ij} = \frac{1}{\sigma_{ij}^2}$) to get the estimated parameters. At this point, we can compute analytically asymptotic estimates of the parameters $\hat{\theta}$ and asymptotic confidence intervals. In this scope, we assume that we are in the case where we have so much observations that the deviation $\Delta\theta$ between the real θ_0 and estimated parameters $\hat{\theta}$ is small. Thus, we can expand the observation function in a Taylor series:

$$y_j^M(t_i, \theta, u) = y_j^M(t_i, \theta_0, u) + \nabla_{\theta} y_j|_{t_i, \theta_0} (\theta - \theta_0). \quad (10)$$

We insert this result in the function to minimize, and we get:

$$\chi^2(\theta) = \sum_{i=1}^N \sum_{j=1}^{L_i} \left[\frac{\epsilon_{ij}^2}{\sigma_{ij}^2} - 2 \frac{\epsilon_{ij}}{\sigma_{ij}^2} \nabla_{\theta} y_j|_{t_i, \theta_0} \Delta\theta + \frac{1}{\sigma_{ij}^2} \Delta\theta^T ([\nabla_{\theta} y_j]^T [\nabla_{\theta} y_j])|_{t_i, \theta_0} \Delta\theta \right]. \quad (11)$$

To minimize $\chi^2(\theta)$, we need to solve the following equation: $\frac{\partial}{\partial \theta} \chi^2(\theta) = 0$, so we get the estimated parameters:

$$\Delta\theta = F^{-1} \sum_{i=1}^N \sum_{j=1}^{L_i} \frac{\epsilon_{ij}}{\sigma_{ij}^2} ([\nabla_{\theta} y_j]^T)|_{t_i, \theta_0}. \quad (12)$$

where F is the Fisher information matrix.

Parameter Estimation and Covariance Matrix. From the knowledge of F we can easily get the exact values of the (asymptotic) estimated parameters:

$$\hat{\theta} = \theta_0 + F^{-1} \sum_{i=1}^N \sum_{j=1}^{L_i} \frac{\epsilon_{ij}}{\sigma_{ij}^2} ([\nabla_{\theta} y_j]^T)|_{t_i, \theta_0}. \quad (13)$$

As we assumed that the residuals are independently distributed, the covariance matrix of the estimated parameter vector is computed by (where the average is over the repetition of experiments):

$$\Sigma = \langle \Delta\theta \Delta\theta^T \rangle = F^{-1}. \quad (14)$$

Thanks to this covariance matrix, we can see the correlation between the parameters. The correlation matrix is defined by:

$$\begin{cases} R_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii} \Sigma_{jj}}}, & \text{if } i \neq j. \\ R_{ij} = 1, & \text{if } i = j. \end{cases} \quad (15)$$

2.2 Parameter Correlation and Identifiability Criteria

Equipped with Eqn. (15), we can measure the interrelationship between the parameters and get an idea of the compensation effects of changes in the parameter

values on the model output. For instance, if two parameters are highly correlated, a change in the model output caused by a change in a model parameter can be compensated by an appropriate change in the other parameter value. This prevents such parameters from being uniquely identifiable even if the model output is very sensitive to changes to individual parameters.

Then we can try to improve the information contained in the data by optimizing one of the criteria derived from Σ . We used the modified E-optimal design: $\min(\frac{\lambda_{max}(\Sigma)}{\lambda_{min}(\Sigma)})$. As it minimizes the ratio of the largest to the smallest eigenvalue, it optimizes the functional shape of the confidence intervals. All calculations have been performed with ByoDyn (<http://cbb1.imim.es/ByoDyn>) most of them at the QosCosGrid [7] environment.

3 Results

3.1 The ODEs Model for the Notch-Delta System

In the model, two adjacent cells, i and j , initially expressing the the same amount of the genes *notch* and *delta* generate an asymmetric final expression of the genes by the lateral inhibition mechanism. The interaction of the protein NOTCH with its ligand Delta activates the cleavage of the NOTCH intracellular domain (NICD) by a γ -secretase. NICD activates the expression of *hes5* and ultimately downregulates *delta*. If one assumes in a very rough approximation that the quantities of the different species in the system are large enough to work with concentrations, we have formalized the verbal model represented in Figure 1, after adimensionalization by $t = T_0 \tau$ and $[x]_i = [x]_0 x_i(\tau)$:

$$\begin{aligned}
 \frac{dnotch_i(\tau)}{d\tau} &= T_0 K_{deg}^{notch} (r_{notch} - notch_i(\tau)) . \\
 \frac{dNOTCH_i(\tau)}{d\tau} &= T_0 K_{deg}^{NOTCH} (notch_i(\tau) - NOTCH_i(\tau)) \\
 &\quad - \frac{k}{n^2} K_{bind}^{ND} [DELTA]_0 NOTCH_i(\tau) \sum_{j=1}^k DELTA_j(\tau) . \\
 \frac{ddelta_i(\tau)}{d\tau} &= T_0 K_{deg}^{delta} \left(1 - \frac{HES5_i^s(\tau)}{\kappa_{HES5} + HES5_i^s(\tau)} - delta_i(\tau) \right) . \\
 \frac{dDELTA_i(\tau)}{d\tau} &= T_0 K_{deg}^{DELTA} (delta_i(\tau) - DELTA_i(\tau)) \\
 &\quad - \frac{k}{n^2} K_{bind}^{ND} [NOTCH]_0 DELTA_i(\tau) \sum_{j=1}^k NOTCH_j(\tau) . \\
 \frac{dND_i(\tau)}{d\tau} &= T_0 \left(\frac{k}{n^2} K_{bind}^{ND} [DELTA]_0 NOTCH_i(\tau) \sum_{j=1}^k DELTA_j(\tau) - K_{deg}^{ND} ND_i(\tau) \right) . \\
 \frac{dhes5_i(\tau)}{d\tau} &= T_0 K_{deg}^{hes5} \left(\frac{ND_i^m(\tau)}{\kappa_{ND} + ND_i^m(\tau)} - hes5_i(\tau) \right) . \\
 \frac{dHES5_i(\tau)}{d\tau} &= T_0 K_{deg}^{HES5} (hes5_i(\tau) - HES5_i(\tau)) . \tag{16}
 \end{aligned}$$

where we have assumed that the NOTCH cleavage after the formation of the ND complex, the NICD transport to the nucleus and the transcription factor activation can be simply approximated by the amount of ND complex that is formed

Fig. 1. Simplified model for the Notch/Delta pathway for two adjacent cells i and j . NOTCH* and DELTA* refers to the activated forms.

on the membrane surface. In 16 *notch* is constitutively activated, while sigmoidal activation and inhibition curves are used for *hes5* and *delta*, respectively.

By using the parameters $K_{deg}^{hes5} = K_{deg}^{HES5} = K_{deg}^{NOTCH} = K_{deg}^{DELTA} = K_{deg}^{ND} = K_{deg}^{delta} = 0.01$; $s = m = 2.0$; $\kappa_{ND} = \kappa_{HES5} = 0.1$; $NOTCH_0 = 5.0$; $K_{deg}^{notch} = 0.0016649$; $K_{bind}^{ND} = 0.25$; $DELTA_0 = 3.0$; $r_{notch} = 0.620926$, the steady state concentrations of the three genes in our model acquire the characteristic chess-like pattern represented in Figure 2, in which different cell types are clearly defined. In addition, the figure shows the correlation matrices for the diverse systems. It appears that the boundary effect vanishes with bigger tissue sizes and that the 5×5 cells model can be considered converged for the purposes of this paper, as seen from the invariant correlation matrix when comparing the 5×5 and 7×7 systems. Thus, in the following paragraphs we will present our protocol for experimental design based on the 5×5 tissue model.

Next, we consider a typical experimental setting in which RT-PCR experiments are carried out and provide time dependent data for each of the genes involved in our model. We will generate hypothetical data from real experiments of inner ear early development in chick[8]. In a typical scenario of the model, 4 tissue samples may be extracted at different stages of development. For each of them RT-PCR experiments may be performed, using three replicas for security, showing a behavior that in the best case will be just close to the simulated concentration profiles from Eq. 16. The parameter set θ may then be globally optimized with several methods. We use here a simple approach consisting on local optimizations from 10 or 100 starting random values of θ with varying value of σ^2 for the generated data points.

Once the fitting parameters are obtained to some approximation, by using the above detailed simple approach or by more sophisticated methods[9], we are interested in improving their confidence intervals. This can be achieved, of course,

Fig. 2. (a.) Delta steady state concentration distribution in model tissues of different dimensions. (b.) Correlation matrices of the adimensional parameters in each case in (a). The steady state is achieved at a biologically plausible time scale.

by choosing a better optimization algorithm or, complementarily, by using information theory in order to estimate what data will provide more information to improve the parameters practical identifiability. This is extremely relevant as new experiments can consume an important number of resources and even one may decide they are not worth trying because of intrinsic identifiability problems of the model. In order to learn about the information content of new data, we generate *in silico* data at 200 time points through the total simulation time t_{total} using the parameters optimized in the previous step. We call this set θ' . In a leave-one-out fashion, each value is deleted at a time and the modified E-criteria is evaluated for the remaining data, in order to discover the computer-generated point, according to the current model (topology plus parameters) that contain more information. Figure 3 shows the result of this approach for the two genes of the system.

In the first iteration of the protocol, the modified E-criteria suggests that new values for the concentration of *hes5* at time $t = 1260$ would be the most informative. At this stage we *measure* new data for that gene at such time step and we proceed the next iteration of the approach again. Such measurement in a real experimental set up is simulated here by a new *in silico* value obtained with or without noise with respect to the known model. Finally, Figure 4 shows the evolution of the modified E-criteria for a number of iterations of the protocol. It can be seen how the higher information content of the new experimental data set (increased after each OED iteration) does not necessarily involves a better (lower) value for the modified E criteria. This problem has multiple origins, being the noise of the new data measured or the fact that the optimization method does not find the same minimum in each parameter estimation step.

Fig. 3. Modified E-criteria after adding one time point per gene at each time from a set of previous targeted behavior

Fig. 4. Evolution of the modified E-criteria for 5 iterations of the OED procedure

4 Conclusions

Optimal experimental design has been demonstrated in a realistic example of *experiment/theory* iterative protocol. In this paper, the *experimental* data is

indeed estimated from new calculations in order to show the general applicability of the protocol, although its migration to real experimental setups is straightforward. The benefits from using the proposed approach are clear, as the new experiments to be carried out are decided from a predicted behavior of the modified E-criteria for a set of *in silico* generated data from the model from optimal parameters from the previous step in the iteration. The proposed protocol provides an easy and neat method to incorporate experimental data, that may be difficult or expensive to obtain, in an informed way. At the same time it provides clues about the identifiability of the parameters for the proposed model, according to the evolution of the modified E-criteria with the iterations of the OED approach. Thus, one expects the modified E-criteria to approach the limit of 1 for a perfectly identifiable model if a big number of experiments is performed, while reaching a different limiting value is indicative of the unidentifiability of the model. The protocol has been exemplified on a hypothetical situation in which a simple gene regulatory network includes three genes interacting in a multicellular system. However, the data proposed, its distribution and the error one performs in the experimental evaluations are realistic and match a typical experimental setting. The next step is to apply this protocol to real data on a more complex model like the regionalization of cellular systems during vertebrate development[10].

Finally, the practical implementation of the protocol makes it suitable for parallelization in several points, like the multiple optimization in each step or the evaluation of the modified E-criteria itself for several time/species trial values.

Acknowledgments. ALGL thanks Generalitat de Catalunya for a PhD fellowship. Partially funded by grant BQU2003-04448 (MCYT: Spanish Ministry of Science and Technology), and EC-STREP projects QosCosGrid (FP6-IST-2005-033883) and BioBridge (FP6-LIFESCIHEALTH-2005-037909). The authors thankfully acknowledge the computer resources and assistance provided by the Barcelona Supercomputing Center.

References

1. Tomline, C.J., Axelrod, J.D.: Biology by numbers: mathematical modelling in developmental biology. *Nature Reviews* 8, 331–340 (2007)
2. Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K.N., Manu, M.E., Vanario-Alonso, C., Samsonova, M., Sharp, D.H., Reinitz, J.: Dynamic control of positional information in the early *Drosophila* embryo. *Nature* 430, 368–371 (2004)
3. de Jong, H.: Modeling and simulation of genetic regulatory systems: a literature review. *J Comput. Biol.* 9(1), 67–103 (2002)
4. Meinhardt, H.: Computational modelling of epithelial patterning. *Curr. Opin. Genet. Dev.* 17(4), 272–280 (2007)
5. von Dassow, G., Meir, E., Munro, E.M., Odell, G.M.: The segment polarity network is a robust developmental module. *Nature* 406, 188–192 (2000)
6. Faller, D., Klingmüller, U.T.J.: Simulation methods for optimal experimental design in systems biology. *Simulation* 79, 717–725 (2003)

7. Coti, C., Herault, T., Peyronnet, S., Rezmerita, A., Cappello, F.: Grid services for MPI. In: ACM/IEEE (ed.) Proceedings of the 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2008), Lyon, France (May 2008)
8. Alsina, B., Abello, G., Ulloa, E., Henriq, D., Pujades, C., Giraldez, F.: FGF signaling is required for determination of otic neuroblasts in the chick embryo. *Dev. Biol.* 267(1), 119–134 (2004)
9. Rodriguez-Fernandez, M., Egea, J.A., Banga, J.R.: Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics* 7, 483 (2006)
10. Alsina, B., Garcia de Lomana, A., Villà-Freixa, F., Giraldez, F.: (submitted, 2008)